**Automatic Image Captioning Based On Real Time Data**

**Group 4**

**Team Members**

Chiru Toleti

Satyam Kumar

Praveen

Yaswanth Kumar

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## INDEX

[**INDEX**](#_1sdu58hn9qfh) **1**

[**Problem Statement**](#_ttwk9h1n9igt) **2**

[**Literature Review**](#_1dnoed517lak) **2**

[**Methodologies**](#_727f2lwk6gs9) **3**

[**Finalised Approach**](#_fj66n1qrunrq) **4**

[**Possible outcomes in stages**](#_o96twxathf11) **6**

[**Challenges**](#_4to0djvicvnv) **6**

[**Applicability in the real world**](#_d0tclwec3qrp) **7**

[**Reference Papers**](#_9lqe3zbr9tqm) **8**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Problem Statement

With the rapid development of digitalization, there are a huge amount of images, accompanied with a lot of related texts. Although it is an easy task for a human to describe an image, it becomes very difficult for a machine to perform such a task. Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. The objective of automatic image captioning is to generate properly formed English sentences to describe the content of an image automatically, which is of great impact in various domains.

## Literature Review

Automatic image captioning has recently attracted much research interest. One common framework used in this category is the encoder-decoder framework for image captioning. Recent research works have used the deep convolutional neural network (CNN) as the encoder and the deep recurrent neural network (RNN) as the decoder. However, it still remains challenging to identify the proper CNN and RNN models for the image captioning.In this paper, we investigate one single-joint mode, AICRL, for automatic image generation using ResNet50 (a convolutional neural network) and LSTM (long short-term memory) with soft attention mechanism. AICRL consists of

1. An encoder and a decoder.We adopt ResNet50 as the encoder to create an extensive representation of an input image by embedding it into a vector.
2. The LSTM with a soft attention as the decoder which selectively focuses the attention over a certain part of an image to predict the next sentences.

Furthermore, we conduct extensive experiments and empirically determine the structure of the model and fine-tuned the model hyperparameters. Our experimental evaluation indicates that AICRL is effective to generate proper captions for the image.

## Methodologies

Much research effort has been devoted to automatic image captioning, and it can be categorised into

1. Template based image captioning
2. Retrieval-based image captioning and
3. Novel image caption generation

### Template-based image captioning:

First detects the objects/attributes/actions and then fills the blanks slots in a fixed template

#### Pros:

The template-based approach is aimed at generating captions by using fixed templates with a number of blank slots, in which way different objects, attributes, and actions are detected first and then the blank spaces in the templates are filled. Template-based methods can generate grammatically correct captions.

#### Cons:

However, templates are predefined and length of captions cannot be variable.

### Retrieval-based image captioning:

Approaches first find the visually similar images with their captions from the training dataset, and then the image caption is selected from similar images with captions.

#### Pros:

The retrieval-based approach tries to generate description for an image by selecting the most semantically similar sentences from the sentence pool or directly copying sentences from other visually similar images.

### Novel image captioning:

Generation approaches are to analyse the visual content of the image and then to generate image captions from the visual content using a **language model**.

Compared to the first two categories, novel caption generation can generate new captions for a given image that are semantically more accurate than previous approaches. Most of the works in this category rely on machine learning and deep learning.

## Finalised Approach

### AICRL - Automatic Image Captioning based on ResNet50 and LSTM with software attention

The ultimate purpose of AICRL is to generate the proper description for the given images. To do so, the AICRL model is designed with an encoder and decoder architecture based on CNN and RNN. In particular, to extract visual features, we use the ResNet50 network as the encoder to generate a one-dimensional vector representation of the input images.

After that, to generate the description sentences, we adopt the LSTM as the language model for the decoder to decode the vector into a sentence.

Meanwhile, we utilise the soft attention in the decoder to enable the model to selectively focus the attention over a certain part of an image to predict the next sentence better. We conduct extensive experiments, empirically determine the structure of the model, and fine-tune the model hyperparameters. The whole model is fully trainable by using a stochastic gradient descent.

#### Image Feature Extraction:

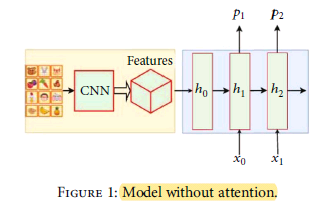
### To represent the image, we adopt the convolutional neural network (CNN), ResNet50.

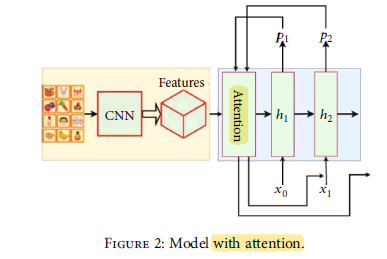
#### The Language Model:

### In our design, we adopt LSTM as our language model to generate proper caption based on the input vector from the ResNet50 output.

#### Attention Mechanism:

Soft attention is implemented by adding an additional input of attention gate into LSTM that helps to concentrate selective attention. The main drawback of the model without attention is that it tries to decode the full image from the last hidden layer of h0 in Figure 1. It is like an analogy with machine translation in the whole process. To do a translation of the whole text is just from the “last word.” So it will lose a lot of useful information from the beginning of the text.





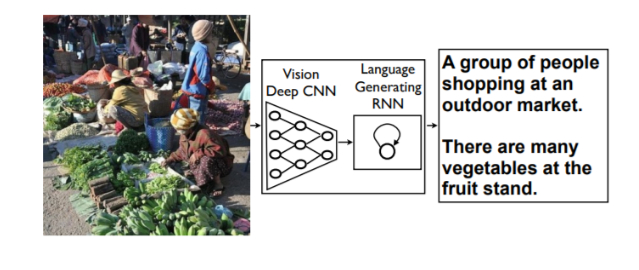
### DataSets can be used:

Dataset of images with their five descriptions for each image.

| **Name of the dataset** | **Training Samples** | **Testing Samples** |
| --- | --- | --- |
| MS COCO 2014 | 102,739 | 20,548 |
| Flickr8K | 7,000 | 1,000 |

## Possible outcomes in stages

CNNs can produce a rich representation of the input image by embedding it to a fixed-length vector, such that this representation can be used for a variety of vision tasks. Hence, it is natural to use a CNN as an image “encoder”, by first pre-training it for an image classification task and using the last hidden layer as an input to the RNN decoder that generates sentences. Which will be the neutral way of Image captioning.



## Challenges

The main drawbacks of the work are the quick model overfitting, so they use the heavy and expensive GoogLeNet.

ResNet50, which is a very deep network that has 50 layers. The depth of the network is crucial for neural networks, but deeper networks are more difficult to train. The structure of ResNet50 facilitates the training of networks and allows them to be much deeper, which leads to increased performance in different tasks. ResNet50 is much deeper than their “simple” counterparts, but moreover, the number of parameters (weights) of such networks is much smaller.

Deep convolutional neural networks have led to a series of breakthroughs for image classification. Recent evidence reveals that network depth is of crucial importance. Many other nontrivial visual recognition tasks have also greatly benefited from the deep models. With the network depth increasing, the accuracy of networks increases rapidly, which is not surprising and then rapidly degrades (saturated). This degradation is not caused by overfitting, and the addition of even more layers leads to a higher learning error. In a sense, this is strange, since a deeper network has a strictly large representational power. It is possible for ResNet50 to get a deeper model trivially, which is not worse than the less deep network. It can be done by adding several identity layers, that is, levels that simply skip the signal further without changes. ResNet50’s deeper levels have to predict the difference between the output of the previous layers and the objective function. They could always drive the weights to 0 and simply skip the signal. Hence, deep residual learning is a good method that makes the network learn to predict deviations from past layers.

To better isolate the image content, we adopt the soft attention mechanism, which has been widely used to solve the problem of image classification, as there is no need to process all pixels of an image. For example, in the classification problem, the background usually plays an insignificant role. Nevertheless, convolutional neural networks, which are the most popular method for solving such a problem, spend the same amount of computational resources on all parts of the image.

## Applicability in the real world

Image captioning has various applications such as recommendations in editing applications, usage in virtual assistants, for image indexing, for visually impaired persons, for social media, logo identification with the help of AI, Image tagging for Ecommerce, photo-sharing service and online catalog and several other natural language processing applications.

## Reference Papers

1. Image Captioning Based on Deep Neural Networks
2. Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention
3. Show and Tell: A Neural Image Caption Generator